The TMF and its Metadata and Linked Data Forum

15th International Open Forum on Metadata

Matthias Löbe[1], Johannes Drepper[2], Sylvie MN Ngouongo[3], Jürgen Stausberg[3], Philippe Verplancke[4]

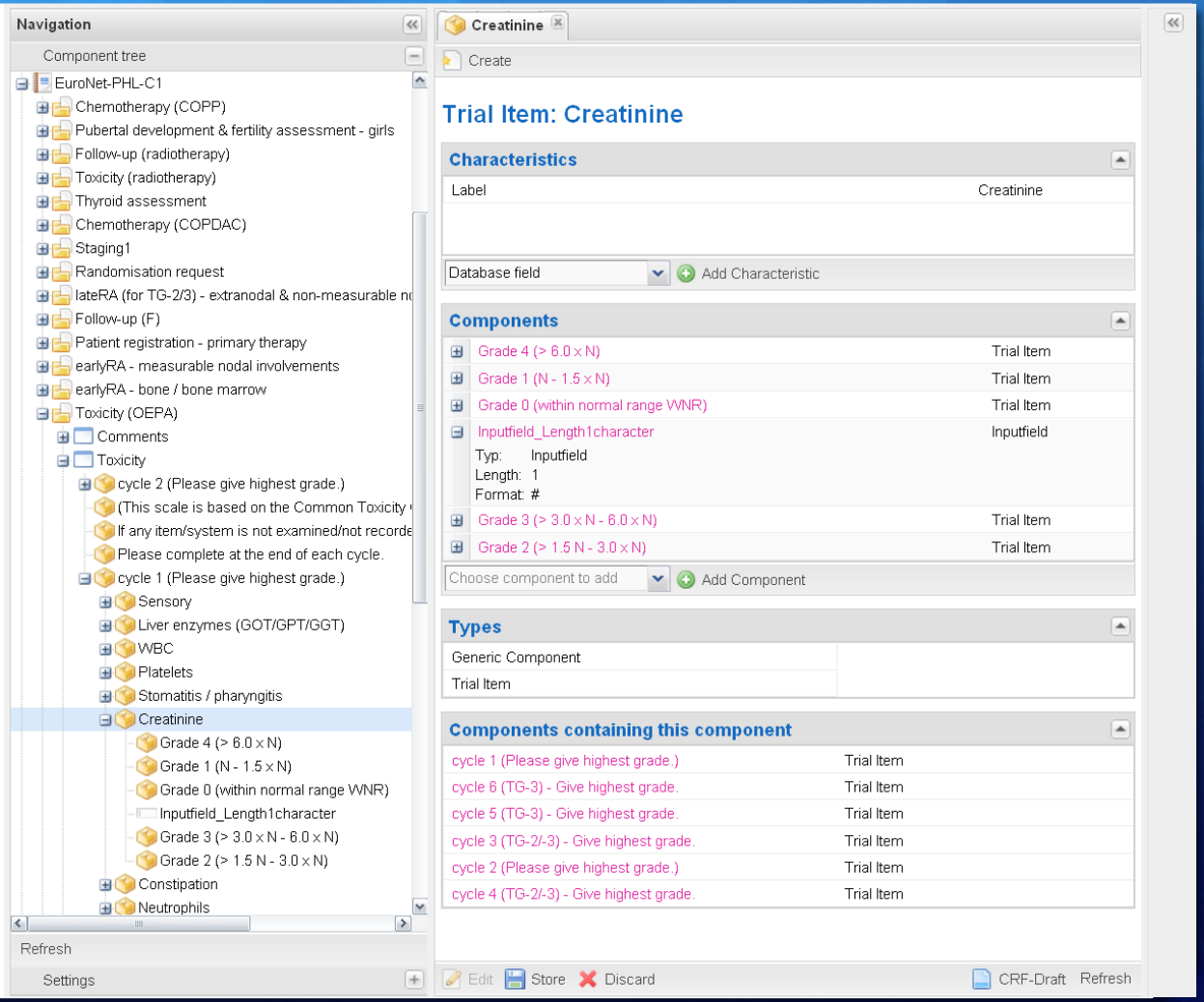
[1] Institute for Medical Informatics (IMISE), Universität Leipzig
[2] TMF - Technology, Methods, and Infrastructure for Networked Medical Research, Berlin
[3] IBE, Ludwig-Maximilians-Universität München
[4] XClinical GmbH, München

Overview

- Introduction
- Current projects in the Telematics Platform with regard to metadata
- The Forum Metadata and Linked Data
- Implications for the application of ISO 11179-3 edition 3 in a community-oriented approach

Starting point: Trial Item Manager (2008)

- RDF/OWL-based tool for repre-senting items in clinical trials
- Planned as an in-house solution only
- Ontological data model was simple but easily extensible



What we intended to replace …

- A poor man‘s trial specification

Vorbekannte Tumorerkrankungen, ggfs mit ICD-10: Suchmöglichkeit ICD-10 C00-D48

Chemotherapie in Anamnese: J/N/weiß nicht

Strahlentherapie in Anamnese: J/N/weiß nicht

[Sonstige Tumorbehandlungen: J/N, bei J Freitext zum Spezifizieren !zunächst rauslassen!]

Relevante Begleiterkrankungen: s. Datei „relevante Begleiterkrankungen“ (der minom Organe → Toxizität)

[Tumor-Histologie n. ICD 0.3: Freitext mit genauer Bezeichnung] ~~zunächst ersetzen durch:~~

Art des Primärtumors

Tumor-Grading: G 1-4

Rezeptorstatus, ~~HER Status~~ HER2-Status

[Patholog. Institut + Einsendenummer)???] zunächst rauslassen

Klin. Bzw. Pathologischer TNM-Status: Pulldown für T - N - M

Source: G. Antony – Workshop ITSM – 09.09.10

We had built it, why didn't they come?

- We'll never know fur sure, but there is some indication …
  - (The software wasn't self-explaining to use and had english labels)
  - (Data managers were not familiar with the terminology used)
  - When there was more content, data managers had no idea how to decide which data elements where superior to others
  - The problem addressed was only a brick in an integrated solution
    - SOPs for data management require CRFs to conform to an Excel template and a certain powerpoint layout
    - Changes made later during database setup were not synchronized
    - Re-use was limited, because in a specific trial, question texts and validation rules are very special
  - No political support („It was always done this way…")
    - Community had too few active contributers
    - Few community features available

Existing Clinical Metadata Repositories

Tool	11179	Content	Tools	API	Community	Open Source
caDSR (USA)	V2	✓	✓	✓	●	●
UK Cancergrid (GB)	V2	?	✓	✓	○	✓
METeOR (Australien)	V2	✓	●	✓	○	○
CIHI (Kanada)	V2	✓	*Currently not available*			
USHIK (USA)	V2	✓	●	○	●	○
MDR (Deutschland)	**V3**	✓★	✓★	✓★	✓★	✓★

Web-based Repository - Software as a Service - Community Approach - Bottom-up-Harmonization

★Planned

Metadata Repository for Clinical and Epimediological Research (MDR)

- Establisment of a national service for providing harmonized data elements and Case Report Forms
  - Based on draft edition 3 of ISO 11179
  - Foundation in Top-Level-Ontology GFO
  - Import for ODM and ClaML files
  - Bottom-up community approach
  - GWT prototype expected in September
- Reference to metadata:
  - obvious

Top TMF Projects Utilizing Metadata

- MDR
- EHR4CR
- KISREK
- Biobank Registry/ P2B2
- ID-Tools
- e-Archiving
- DRT
- Cloud4Health

HIS-based Support for Patient Recruitment for Clinical Trials (KISREK)

- Support for patient recruitment by integrating software tools into the hispotial information system routine workflow (trial registry, query engine, screening list, notification service)
- 5 HIS vendors: Agfa Orbis, Siemens Soarian, Siemens medico, Siemens ISH*med and KAOS
- Reference to metadata:
  - Work package 3: detailed report about the suitability of HIS routine data for recruitment
  - Developed a list of widely-used inclusion/exclusion criteria
  - Decomposition of free-text into „computable criteria" showed:
    - 50% correspond to a single datum in the HIS
    - 30% correspond to two or three dates
  - 50% of all inclusion/exclusion criteria are documented in the HIS in principle
    - But: in many cases incomplete or not in time, decomposition is time-consuming
    - Most suitable: master data, diagnosis, procedures, lab values, observations

Electronic Health Records for Clinical Research (EHR4CR)

- EU project with 33 partners to build a distributed technical platform accessing local data warehouses
- 4 usage scenarios:
  - Protocol feasibility: Leverage clinical data to design viable trial protocols and estimate recruitment (cohort estimation)
  - Patient recruitment: Detect patients eligible for trials to better utilize recruitment potential
  - Clinical trial execution: Re-use routine clinical data to pre-populate trial CRFs
  - Pharmacovigilance: Detect adverse events and collect/transmit relevant information
- Reference to metadata:
  - Development of a central „Pivot Ontology" of 100 data elements for eligibility
  - Semantic mapping from local data to the pivot ontology
  - Local data elements are immutable

Biobank Registry/ P2B2

- National registry for biobanks
- BioMedBridges: EU project providing interoperable services
- Researchers want to maintain control of their data
  - De-centralized peer infrastructure
  - Query tool to request samples
- Reference to metadata:
  - Core Data Set
  - Domain data sets
  - Basic Biobanking Ontology (BBO)

ID Tools

- Data security and privacy are big issues in clinical research
- PID service creates an pseudonym (unique identifier) for a set of patient identification data (similar to HIPAA)
- PSD service creates an second order pseudonym to be managed by a trusted third party
- Reference to metadata:
  - Management of personal data and identifiers
  - Referent tracking

Long-term archiving (LABIMI/F)

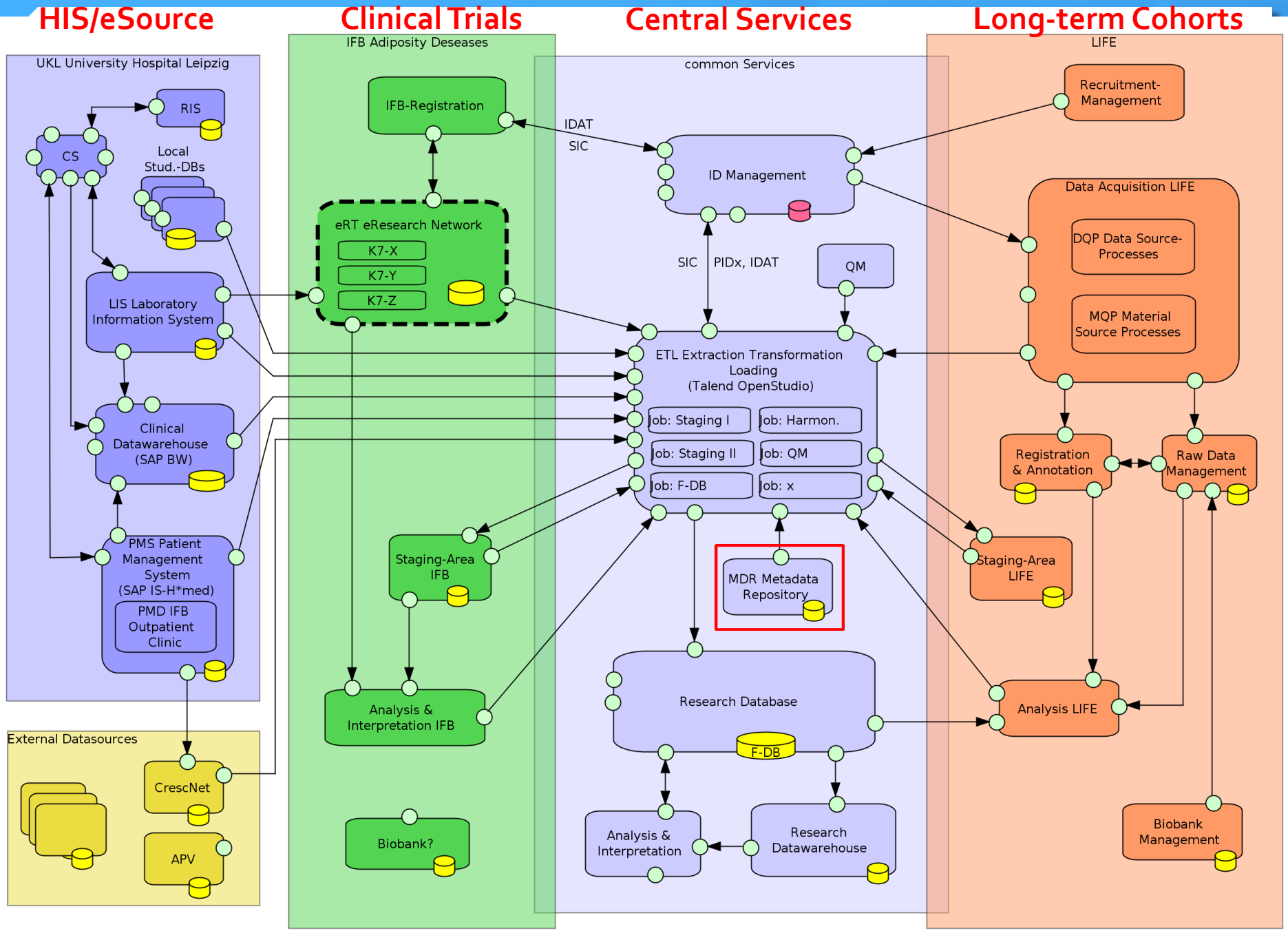
- Archiving of biomedical research data
  - Genomic data
  - Imaging data
- Need for vocabularies to describe!
  - Preservation
  - Provenance
  - Curation
- Reference to metadata:
  - Dublin Core Metadata
  - LOINC, MeSH, SNOMED CT, UMLS

Integrated Data Repository Toolkit (IDRT)

- Provides tools and services around the Harvard i2b2 Data Warehouse software
  - Wizard for semi-automatic installation
  - ETL import jobs for SQL, CSV and ODM files
  - Standard terminologies like ICD-10, OPS, LOINC, MedDRA
  - Data security and privacy via pseudonymization service
- Reference to metadata:
  - Metadata editor to provide mappings and alignments for data elements in the i2b2 ontology cell
  - NCBO BioPortal as ontology source under testing

cloud4health

- Cloud-Computing in Healthcare
  - Secondary Use of unstructured data (text-analysis)
  - Data Warehouse technologies in the cloud
  - Establishment of an infrastructure
- Use Cases:
  - Early detection of adverse events
  - Cost-effectiveness of therapies
- Reference to metadata:
  - Mapping named-entities in discharge letters to SNOMED CT (possibly Observable Entities)


HIS/eSource
Clinical Trials
Central Services
Long-term Cohorts
UKL University Hospital Leipzig
RIS
CS
Local Stud.-DBs
LIS Laboratory Information System
Clinical Datawarehouse (SAP BW)
PMS Patient Management System (SAP IS-H*med)
PMD IFB Outpatient Clinic
External Datasources
CrescNet
APV
IFB Adiposity Deseases
IFB-Registration
eRT eResearch Network
K7-X
K7-Y
K7-Z
Staging-Area IFB
Analysis & Interpretation IFB
Biobank?
common Services
IDAT
SIC
ID Management
SIC
PIDx, IDAT
QM
ETL Extraction Transformation Loading (Talend OpenStudio)
Job: Staging I
Job: Harmon.
Job: Staging II
Job: QM
Job: F-DB
Job: x
MDR Metadata Repository
Research Database
F-DB
Analysis & Interpretation
Research Datawarehouse
LIFE
Recruitment-Management
Data Acquisition LIFE
DQP Data Source-Processes
MQP Material Source Processes
Registration & Annotation
Raw Data Management
Staging-Area LIFE
Analysis LIFE
Biobank Management

Forum Metadata and Linked Data

- Founded June 2011 with a focus on medical research

1. Concepts and Methods:
   - Metadata Models (ISO 11179, CDISC ODM, EN 13606, HL7 CDA)
   - Metadata Artifacts (Std. Values Sets, UCUM, 21090)
   - Metadata Annotations (med. Terminologies, DC, SKOS)
2. Representations und Implementations:
   - Metadata Element Sets (CDASH , HITSP Data Dict., NINDS CDE)
   - Metadata Registries (QA, harmonization, consistency, versioning)
   - Metadata Implementations (Data Integration, Linked Data)

Discussion on 11179

- Is ISO 11179 the Swiss Army Knife?
  - Sophisticated data model
  - More expressive than ODM, Archetypes , CDA
- Some limitations for our use case:
  - Missing features for clinical DM: Order of Data Elements or Value Meanings, repeated occurrences, single choice domains, default values, null values, mandatory fields, cross field checks
  - No classes for modeling document hierarchies or groups of data elements belonging together
  - No composite data elements

Special Challenges for a Community-based Approach

- Users must be able to enter arbitrary data, else the MDR won't attract them
  - What happens to redundant data (**duplicates**)? Which options exist for curating underspecified data elements?
  - Which **user rights**, roles and views are needed and appropriate ?
  - How could **modifications** be tracked and visualized? What implications arise from moving or **deleting metadata items** that are interconnected?
  - How can **harmonization** be supported (reviewed data elements, core data sets)?

Metrics for Excellence

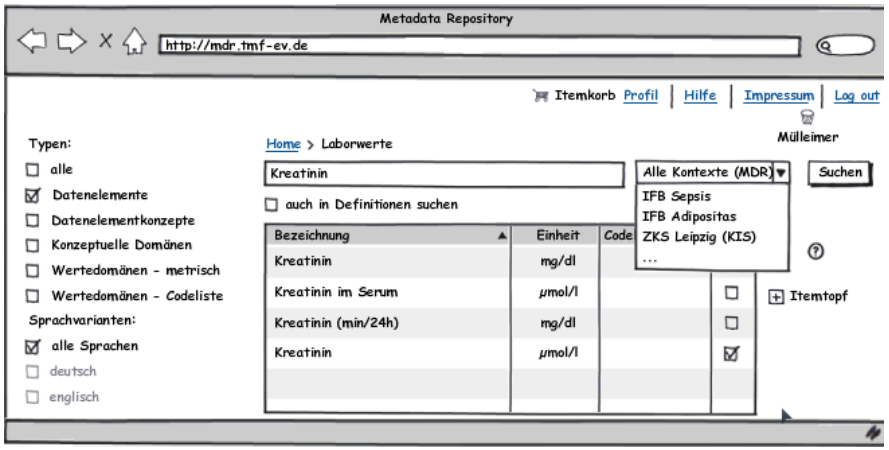
- Quality of the specification
  - Level detail (optional attributes)
  - Consistency (property -> dimensionality -> units)
  - Update frequency (especially if more than one user is involved)
- Rating manually by the creator or the community
  - Adjust to the expertise of the rater
  - Consensus through community voting
- Frequency of Use (for instance in other research projects) – the "common"
  - Adjust to the importance of that project (locally, nationwide, number of subjects)
- Reference to standards:
  - Medical terminologies: ICD, OPS, LOINC, SDTM, SNOMED CT
  - Artifact standards: UCUM measurement units, Null Flavors, ISO 21090 datatypes
  - Contained in Core Data Sets: NINDS CDE, HL7 Value Sets, UK Biobank, …
  - Contained in validated instruments: assessments, scales, scores)

Metrics for Similarity

- What will "Equivalence" or "Similarity" mean with regard to metadata items?
  - Trivial approach: items are equal if there parts are equal
  - Alternatives: items that are conceptually similar
  - Variants: items with different representational values
  - Derivations: items derived by some rule
  - Versions: chronological view on the item's track record
  - Mappings: transformations between data elements
- Most wanted: an ontology for Data Element Concepts
  - Object Classes und Properties as well
- Designations have no influence on similarity

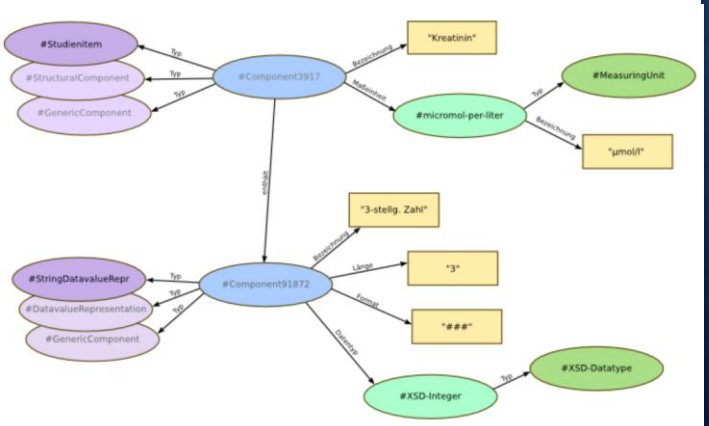
Interfaces for Search and Personalization

- Before one can decide on quality, we need a list of data element candidates
- Currently, that means a textual search in designations, definitions and other text fields
  - Solves morphological problems
  - Problem of synonyms and homonyms persists
  - Import data elements may have misleading or even no designations
  - Components of a data element can have very similar names (data element concept, conceptual domain)
- Facetted search: refinements
  - Metadata objects, kind of usage
  - Research projects, institutions
  - User Profiles, classifications
  - And combination of these



Semantic Web Representation

- Core Data Sets should be part of Linked Data cloud
- We should provide a RDF serialization
- SPARQL endpoint for querying
- Use of Domain Ontologies: OCRe, OBI
- Use of standard vocabularies:
  - FOAF, SKOS , SIOC, SWAN
  - Dublin Core (DC, DCE, DCT)
  - Data Catalog Vocabulary (DCAT)
  - Provenance, Geo, People, Org, Relations



Thank you!